UniHOPE: A Unified Approach for Hand-Only and Hand-Object Pose Estimation

Supplementary Material

In this supplementary material, we provide more qualitative and quantitative results to show the capabilities and robustness of UniHOPE (Sec. A). In Sec. B, we present the implementation details and in Sec. C, we discuss the limitations and future work.

A. More Experimental Results

A.1. Qualitative Results

First of all, we present Figs. A to D, which show that UniHOPE is able to handle both hand-only scenario (left columns) and hand-object scenario (right columns).

Comparison with SOTA Methods. Next, we provide more qualitative comparisons on the DexYCB (Fig. E), HO3D (Fig. F), and FreiHAND datasets (Fig. G).

More De-occluded Examples. Furthermore, we present more de-occluded samples in Fig. H.

A.2. Quantitative Results

Additional Results of Tab. 1. The additional metrics of Tab. 1 in the main paper are provided in Tab. A. Both the metrics before & after PA show an overall performance degeneration of existing HPE/HOPE models when transferring to apply to the other scenario or testing in the original scenario even after re-training on both scenes.

Comparison on Other Splits of DexYCB. We provide the quantitative results of hand pose estimation on the default "S0" split (same distribution for the training and test set) and "S1" split with unseen subjects (train/test: 7/2 subjects) of DexYCB in Tab. B and Tab. C, respectively. Our method achieves the best overall performance, especially in root-relative metrics.

Comparison on HO3D. The remaining hand metrics on HO3D are reported in Tab. D. Though HFL-Net [9] and the combination of H2ONet + HFL-Net achieve better PA results, our method outperforms them by a large margin in the metrics after scale-translation only alignment [4], which takes both the global rotation and hand shape into consideration. We emphasize the importance of global rotation, since it better reflects the visualization quality, as indicated by the qualitative comparison results shown in Fig. F.

A.3. Detailed Analysis on Performance

In this work, we explore a new setting to address HPE and HOPE at once. Applying prior SOTA of HPE/HOPE is suboptimal, even re-trained on all scenarios, as they lack specific designs. For hand-only scenes, HOPE methods are affected by irrelevant object features, even no object is grasped, yet HPE methods may fail for unseen hand poses. For hand-object scenes, HOPE methods lack effective designs to handle severe occlusions, while HPE methods do not utilize object information to enhance performance. Our approach works better in each scene type. As Fig. I shows: (a) when the hand reaches out to grasp an object, our grasp-aware feature fusion reduces the adverse impact of non-grasped object; (b) for unseen hand poses from FreiHAND, our generated de-occluded images introduce richer hand poses to boost performance; (c) our multi-level feature enhancement improves robustness under severe object occlusions; and (d) when grasping objects, our method surpasses HPE methods by leveraging object information. These observations are consistent with the quantitative performance in Tab. 2, 5, 6 in the main paper.

A.4. Additional Ablation Studies

To be consistent with the main paper, we conduct all the ablation studies presented below on DexYCB.

Additional Results of Tab. 7. Since the RHD [22] and Static Gestures Dataset [1] are utilized to fine-tune the ControlNet [12], we also conduct an ablation study of pre-training on these synthetic datasets before training on DexYCB, using a network structure identical to our baseline model with the grasp-aware feature fusion module (Row (b) of Tab. 7 in the main paper). As shown in Tab. F, directly incorporating synthetic datasets into training leads to a minor improvement, indicating the limitation caused by the domain gap between the synthetic and real-world images. Conversely, our occlusion-invariant feature learning strategy substantially enhances the model performance through the foundational data prior provided by ControlNet [20] and the multi-level feature enhancement.

Ablation on Adaptive Control Strength Adjustment. Control strength (ranging from 0 to 1) is imposed on the connections between the ControlNet and Stable Diffusion, controlling the extent to which the output is consistent with the control signal. We propose to adaptively adjust its value with MobRecon [3] pre-trained on DexYCB to avoid tedious manual tuning. The default control strength employed in [12] is 0.55. In our work, we empirically select the candidate control strengths from $\{0.25, 0.4, 0.55, 0.7, 0.85, 1.0\}$, with a similar number of candidates as in [12].

Figure A. UniHOPE is able to handle both hand-only (left column) and hand-object scenarios (right column). Here, we show more qualitative results on DexYCB. For each example, the estimation results are rendered from the original (view 1) and another view (view 2) for clear visualization.

Figure B. More qualitative results of UniHOPE on DexYCB.

Figure C. More qualitative results of UniHOPE across hand-only (left column) and hand-object scenarios (right column) on HO3D.

Figure D. More qualitative results of UniHOPE on HO3D.

HPE	Hand-Only Scene				Hand-Only → Hand-Object Scene				All → Hand-Only Scene				All → Hand-Object Scene			
	J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓	J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓	J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓	J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓
[14]	12.98	5.21	12.52	5.02	19.60 (**-6.62**)	7.71 (**-2.50**)	18.95 (**-6.43**)	7.42 (**-2.40**)	13.16 (**-0.18**)	5.31 (**-0.10**)	12.70 (**-0.18**)	5.11 (**-0.09**)	14.58	6.73	14.10	6.49
[18]	13.34	4.69	13.13	5.05	21.98 (**-8.64**)	7.13 (**-2.44**)	21.42 (**-8.30**)	7.27 (**-2.22**)	14.14 (**-0.80**)	4.74 (**-0.05**)	14.00 (**-0.87**)	5.35 (**-0.30**)	15.20	6.35	15.03	6.74
[21]	14.05	5.55	13.51	5.31	18.37 (**-4.32**)	7.42 (**-1.87**)	17.54 (**-4.03**)	6.91 (**-1.60**)	14.63 (**-0.58**)	5.62 (**-0.07**)	13.96 (**-0.45**)	5.38 (**-0.07**)	14.88	6.74	14.21	6.45

HOPE	Hand-Object Scene				Hand-Object → Hand-Only Scene				All → Hand-Object Scene				All → Hand-Only Scene			
	J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓	J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓	J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓	J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓
[5]	17.99	7.68	17.57	7.88	25.10 (**-7.11**)	7.62 (**+0.06**)	24.40 (**-6.83**)	7.88 (**-0.00**)	18.79 (**-1.00**)	7.77 (**-0.09**)	18.35 (**-0.78**)	7.94 (**-0.06**)	19.75	7.59	19.26	7.98
[9]	14.61	6.56	14.13	6.33	19.39 (**-4.78**)	5.96 (**+0.60**)	18.61 (**-4.48**)	5.75 (**+0.58**)	14.77 (**-0.16**)	6.64 (**-0.08**)	14.29 (**-0.16**)	6.41 (**-0.08**)	13.61	5.20	13.10	5.01

Table A. Full metrics of Tab.1 in the main paper.

	Methods	All Scenes				Hand-Only Scene				Hand-Object Scene			
		J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓	J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓	J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓
HPE	HandOccNet [14]	13.04	5.85	12.61	5.65	13.42	5.39	12.95	5.20	12.79	6.15	12.39	5.95
	MobRecon [3]	14.34	6.50	13.40	5.74	14.57	5.91	13.74	5.29	14.18	6.88	13.19	6.03
	H2ONet [18]	13.89	**5.38**	13.56	5.52	14.10	**4.84**	13.75	5.02	13.76	**5.73**	13.43	5.84
	SimpleHand [21]	13.66	6.02	13.14	5.78	14.48	5.67	13.95	5.46	13.13	6.24	12.62	5.99
HOPE	Liu *et al.* [11]	14.06	5.75	13.57	5.58	14.87	5.47	14.33	5.30	13.53	5.93	13.08	5.75
	Keypoint Trans. [5]	16.61	6.84	16.21	7.05	18.50	7.03	18.00	7.32	15.39	6.71	15.05	6.88
	HFL-Net [9]	13.02	5.58	<u>12.58</u>	<u>5.39</u>	<u>13.41</u>	5.19	<u>12.92</u>	<u>5.00</u>	12.77	5.84	12.35	**5.64**
Unified	H2ONet† + HFL-Net†	13.08	5.47	12.71	5.43	13.81	<u>4.85</u>	13.50	5.06	<u>12.61</u>	5.87	<u>12.20</u>	<u>5.68</u>
	H2ONet‡ + HFL-Net‡	13.30	<u>5.45</u>	12.91	5.40	14.09	<u>4.85</u>	13.74	5.02	12.79	<u>5.83</u>	12.37	**5.64**
	HandOccNet† + HFL-Net†	13.32	5.73	12.87	5.54	14.40	5.50	13.89	5.30	12.63	5.89	12.22	5.69
	HandOccNet‡ + HFL-Net‡	13.43	5.71	12.97	5.51	14.41	5.49	13.90	5.30	12.80	5.85	12.38	5.65
	UniHOPE (ours)	**12.59**	5.54	**12.17**	**5.36**	**12.84**	5.02	**12.38**	**4.85**	**12.42**	5.88	**12.03**	5.69

Table B. Quantitative comparison on DexYCB “S0” split.

	Methods	All Scenes				Hand-Only scene				Hand-Object Scene			
		J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓	J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓	J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓
HPE	HandOccNet [14]	18.33	6.95	17.70	6.71	19.70	6.01	18.95	5.81	17.57	7.47	17.02	7.21
	MobRecon [3]	18.62	7.18	17.73	6.61	19.36	6.27	18.42	5.75	18.21	7.68	17.36	7.09
	H2ONet [18]	18.40	<u>6.40</u>	17.90	6.57	18.92	**5.44**	18.36	5.70	18.11	6.93	17.64	7.05
	SimpleHand [21]	<u>17.38</u>	6.82	<u>16.81</u>	6.73	18.86	6.02	18.14	5.92	16.57	7.26	16.08	7.17
HOPE	Liu *et al.* [11]	17.82	6.46	17.19	<u>6.25</u>	19.12	5.89	18.36	5.69	17.10	**6.77**	16.54	**6.55**
	Keypoint Trans. [5]	21.61	8.15	21.18	8.36	22.84	7.32	22.24	7.59	20.93	8.61	20.60	8.79
	HFL-Net [9]	17.77	6.58	17.16	6.36	<u>18.42</u>	5.72	<u>17.72</u>	<u>5.52</u>	17.41	7.06	16.86	6.82
Unified	H2ONet† + HFL-Net†	17.49	**6.36**	16.94	<u>6.25</u>	19.24	5.50	18.60	5.59	<u>16.54</u>	<u>6.83</u>	<u>16.02</u>	<u>6.61</u>
	H2ONet‡ + HFL-Net‡	17.96	6.48	17.41	6.42	18.92	<u>5.45</u>	18.35	5.69	17.44	7.05	16.89	6.82
	HandOccNet† + HFL-Net†	17.84	6.53	17.22	6.31	20.18	5.95	19.39	5.75	16.55	6.85	16.03	6.62
	HandOccNet‡ + HFL-Net‡	18.63	6.73	17.99	6.50	20.72	6.11	19.91	5.91	17.48	7.06	16.93	6.83
	UniHOPE (ours)	**16.84**	6.42	**16.25**	**6.20**	**17.80**	5.50	**17.11**	**5.30**	**16.31**	6.93	**15.79**	6.70

Table C. Quantitative comparison on DexYCB “S1” split.

To assess the effectiveness of our adaptive control strength adjustment, we compare our model (Row (c) of Tab. 7 in the main paper) with the ones trained with generated samples under fixed control strengths without incorporating the feature enhancement constraints. As shown in Tab. E, our adaptive strategy achieves the best performance in hand pose estimation compared to several control strengths. The samples generated under all candidate control strengths are provided in Fig. J, showing the need to adaptively select control strength for different cases.

Effects of Hyperparameters. The default value of hyperparameter α is empirically set to 10 in Eq. (11) of the main paper. This is to ensure a prediction accuracy over 95%. For the hyperparameters controlling the feature enhancement at three different levels, we evaluate their effects on the hand pose estimation performance in Tab. G. Since the MANO-level feature is a late-stage feature employed to directly regress the final hand pose, an adaption layer is deployed to improve the knowledge transfer. We set a larger value for γ_{MANO} to aim to strongly enforce this feature adaptation process. In our experiments, the values for γ_{init}, γ_{RoI}, and γ_{MANO} are set to 0.1, 0.1, and 0.5, respectively.

A.5. Computational Cost and Efficiency

The training time of our model is 3 days for DexYCB (376k samples) and 12 hours for HO3D (66k samples), respectively, on eight NVidia RTX 2080Ti GPUs.

Tab. H reports the inference speed (FPS, tested on a sin-

	Methods	*Procrustes Alignment*						*Scale-Translation Aligned*	
		J-PE ↓	J-AUC ↑	V-PE ↓	V-AUC ↑	F@5 ↑	F@15 ↑	J-PE ↓	J-AUC ↑
HPE	HandOccNet [14]	10.26	7.95	10.21	79.61	50.61	94.47	28.18	49.28
	MobRecon [3]	10.47	79.14	10.76	78.54	47.57	93.59	29.36	49.36
	H2ONet [18]	9.52	80.97	9.60	80.81	52.62	95.09	29.67	48.53
	SimpleHand [21]	11.28	77.66	11.58	77.05	45.78	91.74	28.41	49.32
HOPE	Liu *et al.* [11]	9.46	81.12	9.39	81.25	54.93	95.64	28.44	49.79
	Keypoint Trans. [5]	12.00	76.24	12.18	75.83	44.71	91.60	40.00	36.36
	HFL-Net [9]	9.01	82.02	8.92	82.18	57.01	96.19	27.97	51.33
Unified	H2ONet† + HFL-Net†	9.49	81.04	9.43	81.16	54.54	95.54	30.60	48.93
	H2ONet‡ + HFL-Net‡	**8.97**	**82.10**	**8.88**	**82.26**	**57.08**	**96.22**	28.00	51.44
	HandOccNet† + HFL-Net†	9.56	80.89	9.50	81.02	54.23	95.47	30.29	49.09
	HandOccNet‡ + HFL-Net‡	9.05	81.94	8.96	82.10	56.79	96.14	27.83	51.45
	UniHOPE (ours)	9.60	80.82	9.45	81.12	52.57	95.68	**25.53**	**53.70**

Table D. Quantitative comparison (*Procrustes Alignment & Scale-Translation Aligned*) on HO3D.

Control Strength Selection	*Root-relative*		*Procrustes Align.*	
	J-PE ↓	V-PE ↓	J-PE ↓	V-PE ↓
s = 0.4	13.76	13.30	5.85	5.65
s = 0.55	13.51	13.06	5.78	5.57
s = 0.7	13.43	12.98	5.75	5.55
Adaptive Adjustment (ours)	**13.38**	**12.92**	**5.71**	**5.52**

Table E. Quantitative results of our adaptive control strength adjustment *vs*. fixed control strengths.

Models	*Root-relative*		*Procrustes Align.*	
	J-PE ↓	V-PE ↓	J-PE ↓	V-PE ↓
Baseline w/ Grasp-aware Feature Fusion	13.84	13.37	5.79	5.58
w/ RHD [22] & Static Gestures [1]	13.79	13.32	5.73	5.53
Ours	**13.03**	**12.59**	**5.59**	**5.40**

Table F. Comparison with directly training with synthetic datasets used by [12].

γ_{init}/ γ_{RoI} / γ_{MANO}	*Root-relative*		*Procrustes Align.*	
	J-PE ↓	V-PE ↓	J-PE ↓	V-PE ↓
0.001 / 0.001 / 0.005	13.17	12.72	5.61	5.41
0.01 / 0.01 / 0.05	13.13	12.69	5.62	5.42
0.1 / 0.1 / 0.5 (ours)	**13.03**	**12.59**	**5.59**	**5.40**
1.0 / 1.0 / 5.0	13.15	12.70	5.70	5.50
10.0 / 10.0 / 50.0	14.13	13.65	6.08	5.87

Table G. Effects of various hyperparameters of the multi-level feature constraints.

gle NVidia RTX 2080Ti GPU), FLOPs, and number of parameters of various models. Thanks to the lightweight object switcher in UniHOPE, UniHOPE has similar inference efficiency and model complexity as HFL-Net [9]. Compared to other SOTA models, UniHOPE has a moderate model size and running speeds, enabling real-time applications.

B. Implementation Details

Scene Division. Following [19], the thresholds for RRE and RTE in grasping label preparation are 5° and 10mm, respectively. An image is categorized into the hand-only scenes, if determined as non-grasping, otherwise hand-object scenes. The numbers of samples in the two scenes are shown in Tab. I. Note that although FreiHAND [23] contains a small number of images interacting with objects in both training and test sets, it cannot be divided due to the lack of object annotations.

Generative De-occluder. We adopt the officially-released pre-trained weights from [12], which fine-tunes ControlNet with synthetic hand images [1, 22]. The hand-object mask is obtained by applying dilation on the render mask of the 3D hand and object to ensure the hand-object region is covered for repainting. Then, we crop the original input image in the training set centered on the hand-object region and resize it to 512×512. The hand-object image and the hand-object mask are fed into the inpainting Stable Diffusion model, conditioned by the hand depth map. Besides, we adopt the positive prompt "a hand grasping gesture, indoor, in the lab" for image generation from the two laboratory benchmarks [2, 4], and the negative prompt is similar to the one in [12]. During inference, the number of reverse steps for DDIM is set to 50 by default.

Network Structure. (i) **Backbone**: Following [9], we adopt ResNet50 [6] as the backbone to extract features from the input image, in which a dual stream structure is adopted to relieve the competition between hand features and object features. (ii) **Hand Encoder**: The hand encoder takes $\mathbf{F}^{OH}$ as input, first using an hourglass network [13] to regress a feature map and the heatmap of 2D hand joints. Then, they are fused via a convolution layer and an element-wise addition, followed by four residual blocks to yield a 1024-dimensional vector. (iii) **MANO Decoder**: It consists of two fully connected layers to predict the hand pose and

Methods	HandOccNet [14]	MobRecon [3]	H2ONet [18]	SimpleHand [21]	Liu *et al.* [11]	Keypoint Trans. [5]	HFL-Net [9]	H2ONet + HFL-Net	HandOccNet + HFL-Net	Ours
FPS	48	78	62	41	51	33	43	36	30	44
FLOPs	15.48G	0.46G	0.74G	9.96G	39.44G	12.66G	10.01G	0.77G / 10.04G	15.51G / 10.04G	10.04G
# Param.	37.22M	8.23M	25.88M	48.89M	34.48M	52.79M	46.08M	72.26M	83.60M	46.38M

Table H. Efficiency comparison with previous methods. Note that FLOPs for the "A+B" methods depend on the predicted grasping status, therefore reported as "FLOPs of (classifier + A) / FLOPs of (classifier + B)".

Datasets (splits)	*Training Set*			*Test Set*		
	All Scenes	Hand-Only Scene	Hand-Object Scene	All Scenes	Hand-Only Scene	Hand-Object Scene
DexYCB "S0"	401,507	153,210	248,297	78,768	30,848	47,920
DexYCB "S1"	351,943	138,775	213,168	104,128	36,912	67,216
DexYCB "S3"	376,374	145,051	231,323	76,360	29,912	46,448
HO3D	66,034	5,595	60,439	11,524	2,971	8,553
FreiHAND	130,240	N/A	N/A	3,960	N/A	N/A

Table I. Number of samples in hand-only/hand-object scenes for different datasets (splits).

shape parameters of the MANO model from the feature produced by the hand encoder. (iv) **Object Decoder**: Following [9], the feature after RoIAlign from the hand branch is fused with the one from the object branch through a cross-attention layer, to enhance the object feature learning. The fused feature is then forwarded through six convolutional layers to predict the 2D projections of the 3D object corner keypoints and corresponding confidence. In testing, the object pose is computed by the Perspective-n-Point (PnP) algorithm [8] using the correspondence between the predicted 2D and the original 3D keypoints on the object mesh.

Training Details. Following [9], we perform data augmentation on the training samples, including random scaling ($\pm 20\%$), rotating ($\pm 180°$), translating ($\pm 10\%$), and color jittering ($\pm 50\%$). Our training process consists of two stages. In the first stage, the de-occluded images are incorporated into training without the feature enhancement loss for 30 epochs to first adapt the model to the domain of the generated data. In the second stage, the network is additionally supervised by the enhancement constraints between the image pairs for another 40 epochs under the same setting.

C. Limitations and Future Work

Limitations. Though we are able to predict the grasping status of unseen objects, the performance of their pose estimation tends to degrade when the object shape/appearance varies largely, due to the limited object categories in the training data. Besides, despite being provided in most existing public benchmarks, the object annotations are lacking in certain datasets, limiting the applicability of our approach as they are required for scene division and inpainting masks.

Future Work. To improve the model's generalizability towards unseen objects, a promising direction is to utilize the knowledge prior from the various vision foundation models [7, 10, 15], which demonstrated remarkable performance in zero-shot scenarios. Another approach that we are considering for improving the model's generalizability is to train on large-scale synthetic data by leveraging diffusion models [16, 19] or large language models [17].

References

[1] Synthesis AI. Static gestures dataset, data retrieved from synthesis ai. https://synthesis.ai/static-gestures-dataset/, 2023. 1, 7

[2] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. DexYCB: A benchmark for capturing hand grasping of objects. In *CVPR*, pages 9044–9053, 2021. 7

[3] Xingyu Chen, Yufeng Liu, Yajiao Dong, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. MobRecon: Mobile-friendly hand mesh reconstruction from monocular image. In *CVPR*, pages 20544–20554, 2022. 1, 6, 7, 8

[4] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. HOnnotate: A method for 3D annotation of hand and object poses. In *CVPR*, pages 3196–3206, 2020. 1, 7

[5] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3D pose estimation. In *CVPR*, pages 11090–11100, 2022. 6, 7, 8

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 7

[7] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 8

[8] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua.

EPnP: An accurate O(n) solution to the PnP problem. *IJCV*, 81:155–166, 2009. 8

[9] Zhifeng Lin, Changxing Ding, Huan Yao, Zengsheng Kuang, and Shaoli Huang. Harmonious feature learning for interactive hand-object pose estimation. In *CVPR*, pages 12989–12998, 2023. 1, 6, 7, 8

[10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36, 2024. 8

[11] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3D hand-object poses estimation with interactions in time. In *CVPR*, pages 14687–14697, 2021. 6, 7, 8

[12] Wenquan Lu, Yufei Xu, Jing Zhang, Chaoyue Wang, and Dacheng Tao. HandRefiner: Refining malformed hands in generated images by diffusion-based conditional inpainting. In *ACM MM*, 2024. 1, 7

[13] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499. Springer International Publishing, 2016. 7

[14] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. HandOccNet: Occlusion-robust 3D hand mesh estimation network. In *CVPR*, pages 1496–1505, 2022. 6, 7, 8

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 8

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, pages 6000–6010, 2017. 8

[17] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. FoundationPose: Unified 6D pose estimation and tracking of novel objects. In *CVPR*, pages 17868–17879, 2024. 8

[18] Hao Xu, Tianyu Wang, Xiao Tang, and Chi-Wing Fu. H2ONet: Hand-occlusion-and-orientation-aware network for real-time 3D hand mesh reconstruction. In *CVPR*, pages 17048–17058, 2023. 6, 7, 8

[19] Hao Xu, Haipeng Li, Yinqiao Wang, Shuaicheng Liu, and Chi-Wing Fu. HandBooster: Boosting 3D hand-mesh reconstruction by conditional synthesis and sampling of hand-object interactions. In *CVPR*, pages 10159–10169, 2024. 7, 8

[20] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 1

[21] Zhishan Zhou, Shihao Zhou, Zhi Lv, Minqiang Zou, Yao Tang, and Jiajun Liang. A simple baseline for efficient hand mesh reconstruction. In *CVPR*, pages 1367–1376, 2024. 6, 7, 8

[22] Christian Zimmermann and Thomas Brox. Learning to estimate 3D hand pose from single RGB images. In *ICCV*, pages 4903–4911, 2017. 1, 7

[23] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images. In *ICCV*, pages 813–822, 2019. 7

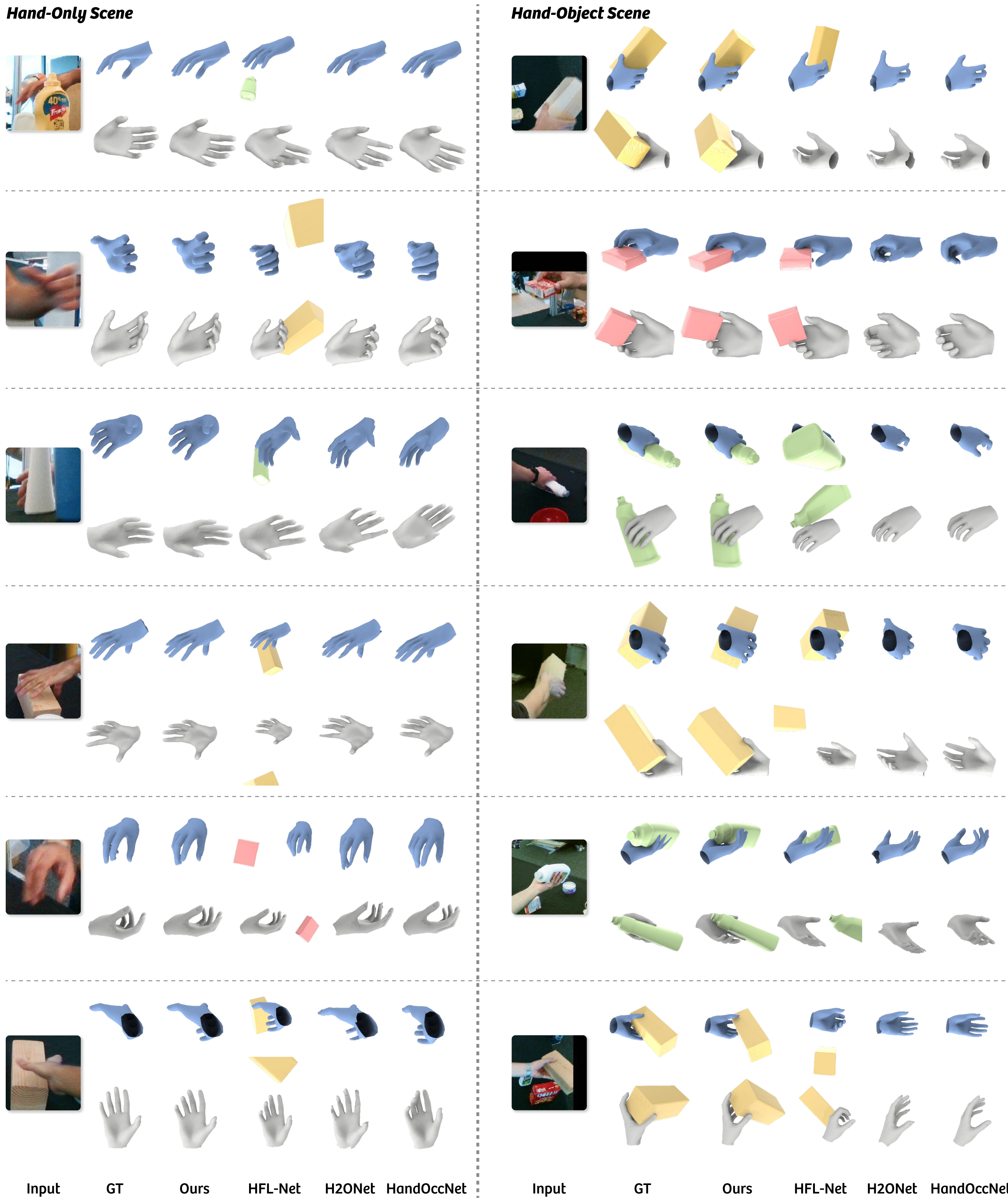

Figure E. Qualitative comparison between UniHOPE and SOTA HPE/HOPE methods across hand-only/hand-object scenarios in DexYCB ("S3" split), in which all the grasping objects are unseen during training.

Figure F. Qualitative comparison between UniHOPE and SOTA HPE/HOPE methods across hand-only/hand-object scenarios in HO3D. The ground truths are not publicly available.

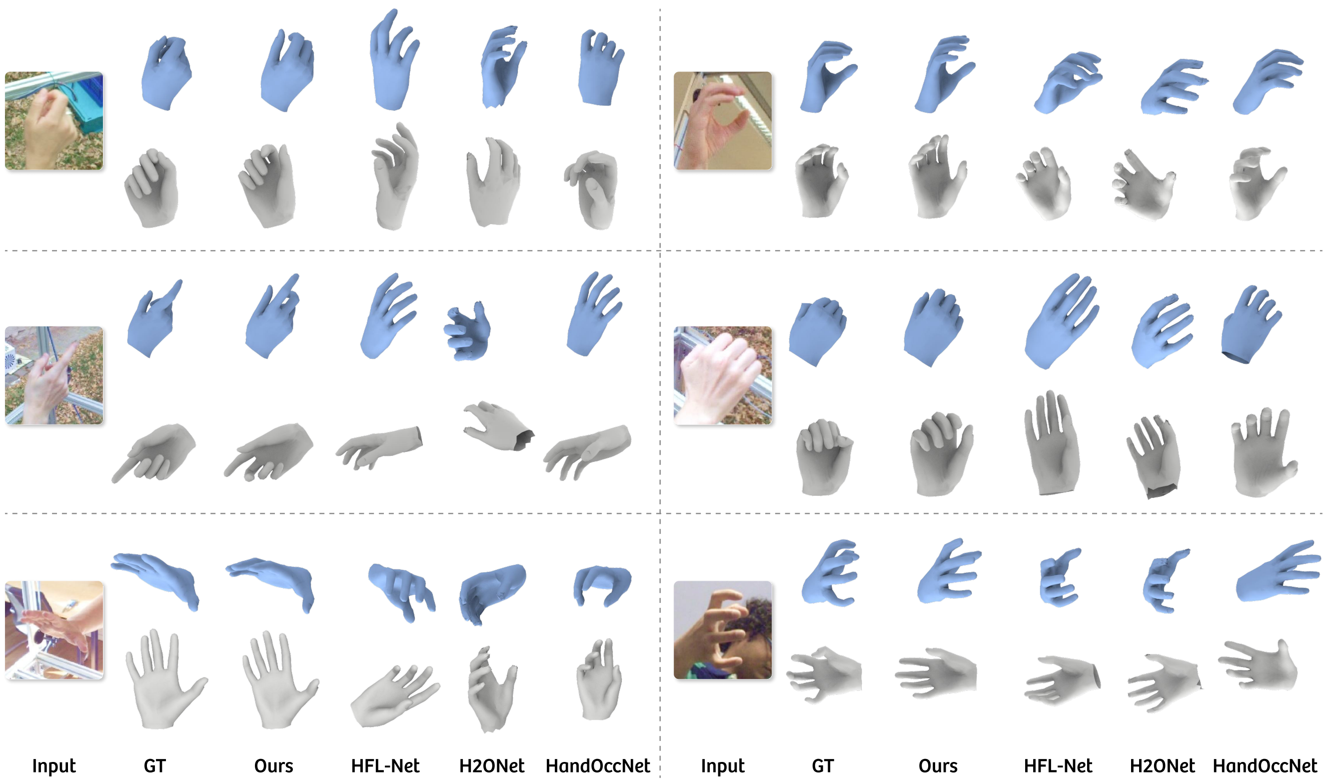

Figure G. Qualitative comparison between our method and SOTA HPE/HOPE methods on FreiHAND.

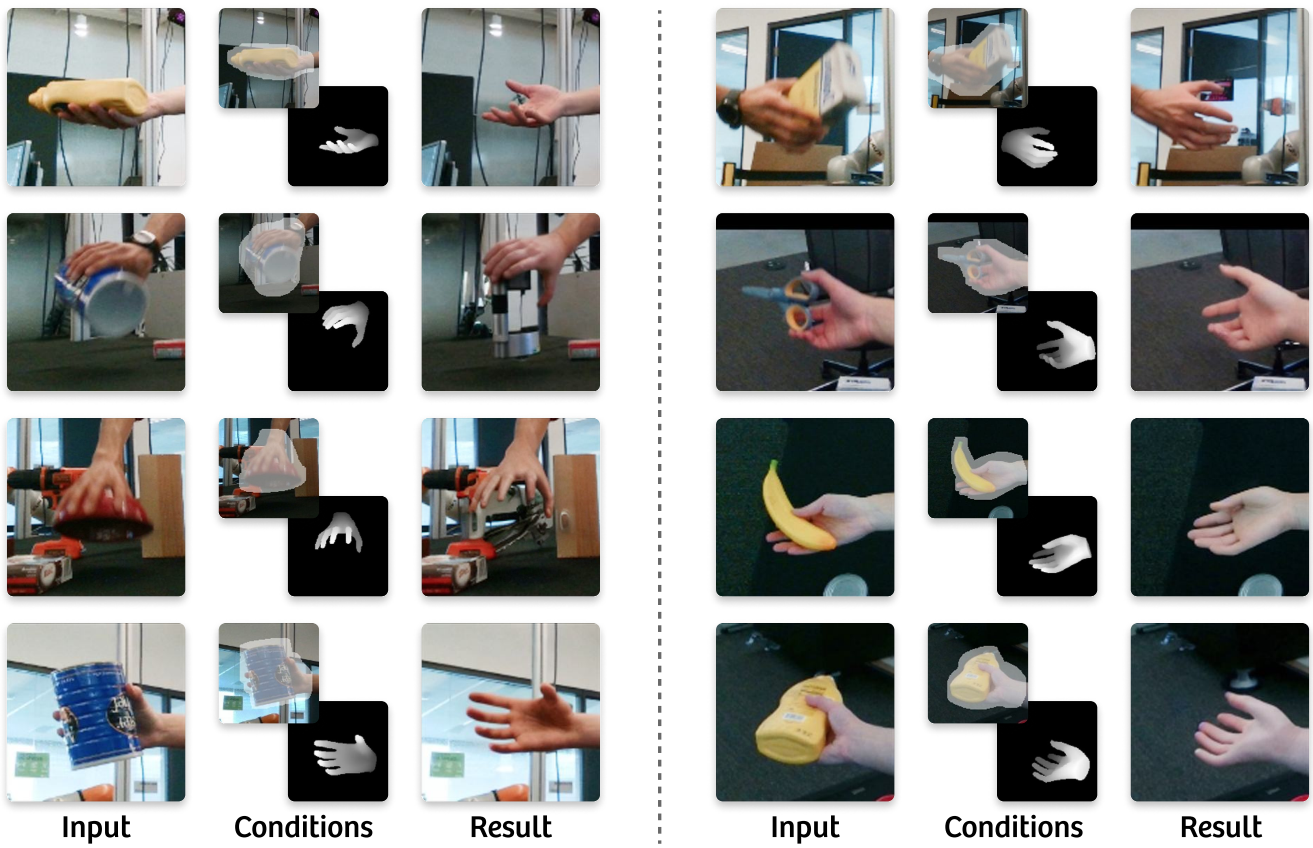

Figure H. More examples of de-occluded hand images. Note that masks are overlaid on the original image for better visualization, the actual condition for our generative de-occluder is a binary mask.

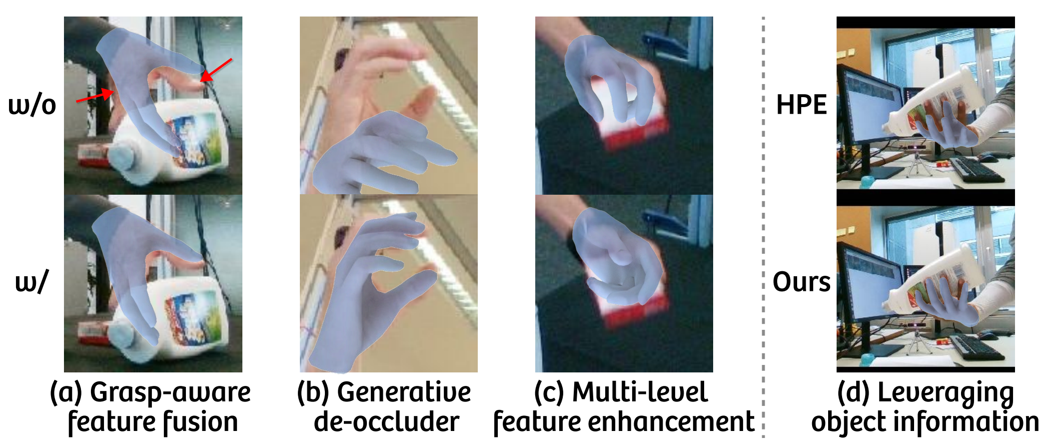

Figure I. Effects of different designs in our pipeline.

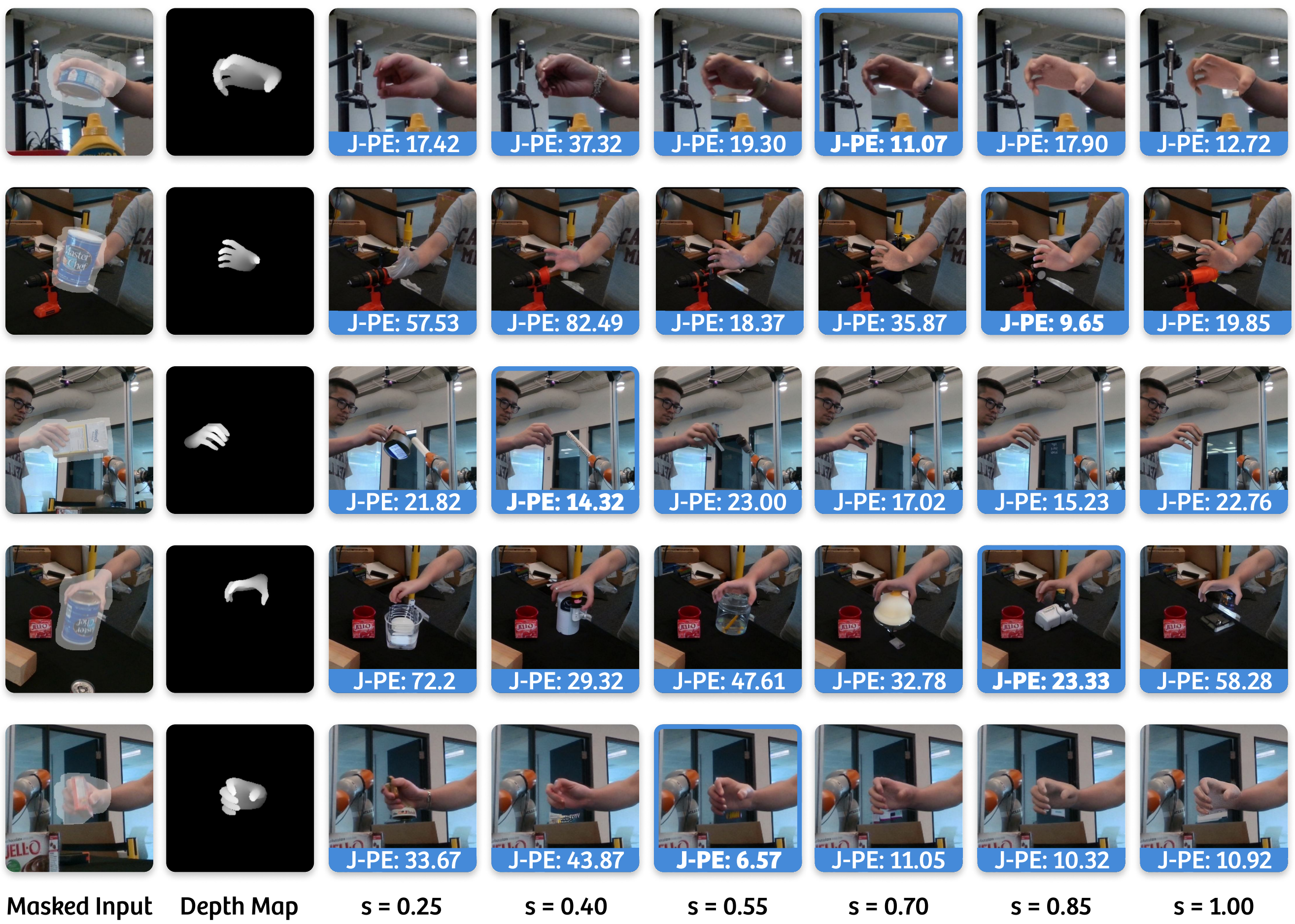

Figure J. The generated images with varying control strengths. Our adaptive strategy (metrics marked in **bold**) effectively balances fidelity and consistency.